

Average path length in random networks

Agata Fronczak, Piotr Fronczak and Janusz A. Hołyst

Faculty of Physics and Center of Excellence for Complex Systems Research,
Warsaw University of Technology, Koszykowa 75, PL-00-662 Warsaw, Poland

(Dated: February 1, 2008)

Analytic solution for the average path length in a large class of random graphs is found. We apply the approach to classical random graphs of Erdős and Rényi (*ER*) and to scale-free networks of Barabási and Albert (*BA*). In both cases our results confirm previous observations: small world behavior in classical random graphs $l_{ER} \sim \ln N$ and ultra small world effect characterizing scale-free *BA* networks $l_{BA} \sim \ln N / \ln \ln N$. In the case of scale-free random graphs with power law degree distributions we observed the saturation of the average path length in the limit of $N \rightarrow \infty$ for systems with the scaling exponent $2 < \alpha < 3$ and the small-world behaviour for systems with $\alpha > 3$.

PACS numbers: 89.75.-k, 02.50.-r, 05.50.+q

During the last few years random, evolving networks have become a very popular research domain among physicists [1, 2, 3, 4]. A lot of efforts were put into investigation of such systems, in order to recognize their structure and to analyze emerging complex properties. It was observed that despite network diversity, most of real web-like systems share three prominent features: small average path length (*APL*), high clustering and skewed degree distribution [1, 2, 3, 4, 5]. Several network topology generators have been proposed to embody the fundamental network characteristics [6, 7, 8, 9]. Due to extensive numerical simulations there were created and analyzed realistic models of real networks especially basing on preferential attachment rule introduced by Barabási and Albert [3, 4, 9]. The most basic issues within the scope of network investigation are structural: connectivity distributions [8, 9, 10], correlation analyzes (including clustering) [11, 12, 13] and finally estimations of the *APL* [14, 15, 16, 17]. The last characteristics is of great importance for network studies as it delivers basic information on a type of network geometry. It is clear, that a better understanding of network topology is of great importance for modern network designing and indirectly affects such crucial fields like information processing in different communication systems (including the Internet) [18, 19, 20, 21], disease or rumor transmission in social networks [22, 23, 24] and network optimization [25, 26, 27]. All these processes become more efficient when the mean distance between network sites is smaller.

It is well known that random networks such as Erdős and Rényi (*ER*) graphs, as well as partially random networks such as Watts-Strogatz *small-world* models [15, 28], have a very small *APL*, which scales as $l \sim \ln N$, where N describes the network size. In fact, it was expected that the logarithmic size effect on the *APL* is a common property of random networks [16]. Very recently Cohen and Havlin found [17] that random networks with power-law degree distribution $P(k) \sim k^{-\alpha}$ and the scaling exponent $2 < \alpha < 3$ exhibit anomalous scaling of the average distance $l \sim \ln \ln N$. Such an anomalous

scaling is expected to lead to anomalies in diffusion and transport phenomena within the networks. The result is particularly interesting since it is known that most of real networks, including both manmade communication networks like the Internet and natural networks like food or metabolic networks, exhibit scale-free character with the relevant scaling exponents [1, 2, 3, 4].

The paper presents an analytic theory describing metric features of random networks. It allows to calculate the main network characteristics like: *APL*, intervertex distance distribution and the mean number of vertices at a certain distance away from a randomly chosen vertex. We compare our analytic results with numerical simulations performed for *ER* random graphs and for scale-free Barabási and Albert (*BA*) networks.

Let us start with the following lemma.

Lemma 1 *If A_1, A_2, \dots, A_n are mutually independent events and their probabilities fulfill relations $\forall_i P(A_i) \leq \varepsilon$ then*

$$P\left(\bigcup_{i=1}^n A_i\right) = 1 - \exp\left(-\sum_{i=1}^n P(A_i)\right) - Q, \quad (1)$$

where $0 \leq Q < \sum_{j=0}^{n+1} (n\varepsilon)^j / j! - (1 + \varepsilon)^n$.

Proof. Using the method of inclusion and exclusion [29] we get

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{j=1}^n (-1)^{j+1} S(j), \quad (2)$$

with

$$\begin{aligned} S(j) &= \sum_{1 \leq i_1 < i_2 < \dots < i_j \leq n} P(A_{i_1}) P(A_{i_2}) \dots P(A_{i_j}) \\ &= \frac{1}{j!} \left(\sum_{i=1}^n P(A_i) \right)^j - Q_j, \end{aligned} \quad (3)$$

where $0 \leq Q_j \leq (n^j / j! - \binom{n}{j}) \varepsilon^j$. The term in bracket represents the total number of redundant components occurring in the last line of (3). Neglecting Q_j it is easy to

see that $(1 - P(\cup A_i))$ corresponds to the first $(n+1)$ terms in the MacLaurin expansion of $\exp(-\sum P(A_i))$. The effect of higher-order terms in this expansion is smaller than $R < (n\varepsilon)^{n+1}/(n+1)!$. It follows that the total error of (1) may be estimated as $Q < \sum_{j=1}^n Q_j + R$. This completes the proof.

Let us notice that the terms Q_j in (3) disappear when one approximates multiple sums $\sum_{1 \leq i_1 < i_2 < \dots < i_j \leq n}$ by corresponding multiple integrals. For $\varepsilon = A/n \ll 1$ the error of the above assessment is less than $A^2 \exp(A)/n$ and may be dropped in the limit $n \rightarrow \infty$.

A random graph with a given degree distribution $P(k)$ is the simplest network model [16]. In such a network the total number of vertices N is fixed. Degrees of all vertices are independent identically distributed random integers drawn from a specified distribution $P(k)$ and there are no vertex-vertex correlations. Because of the lack of correlations the probability that there exists a walk of length x crossing index-linked vertices $\{i, v_1, v_2 \dots v_{(x-1)}, j\}$ is described by the product $\tilde{p}_{iv_1} \tilde{p}_{v_1 v_2} \tilde{p}_{v_2 v_3} \dots \tilde{p}_{v_{(x-1)} j} | v_{(x-2)} v_{(x-1)}$ where

$$\tilde{p}_{ij} = \frac{k_i k_j}{\langle k \rangle N}, \quad (4)$$

gives a connection probability between vertices i and j with degrees k_i and k_j respectively, whereas

$$\tilde{p}_{ij|i} = \frac{(k_i - 1)k_j}{\langle k \rangle N} \quad (5)$$

describes the conditional probability of a link $\{i, j\}$ given that there exists *another* link $\{l, i\}$. It is important to stress that the graph theory distinguishes a *walk* from a *path* [30]. A walk is just a sequence of vertices. The only condition for such a sequence is that two successive nodes are the nearest neighbors. A walk is termed a path if all of its vertices are distinct. In fact we are interested in the shortest paths. Let us consider the situation when there exists at least one walk of the length x between the vertices i and j . If the walk(s) is(are) the shortest path(s) i and j are exactly x -th neighbors otherwise they are closer neighbors. In terms of statistical ensemble of random graphs [31] the probability $p_{ij}(x)$ of at least one walk of the length x between i and j expresses also the probability that these nodes are neighbors of order not higher than x . Thus, the probability that i and j are exactly x -th neighbors is given by the difference

$$p_{ij}^*(x) = p_{ij}(x) - p_{ij}(x-1). \quad (6)$$

In order to write the formula for $p_{ij}(x)$ we take advantage of the lemma (1)

$$p_{ij}(x) = 1 - Q - \exp\left[-\sum_{v_1=1}^N \sum_{v_2=1}^N \dots \sum_{v_{(x-1)}=1}^N \tilde{p}_{iv_1} \dots \tilde{p}_{v_{(x-1)}j} | v_{(x-2)} v_{(x-1)}\right], \quad (7)$$

where N is the total number of vertices in a network. A sequence of $(x+1)$ vertices $\{i, v_1, v_2 \dots, v_{(x-1)}, j\}$ beginning with i and ending with j corresponds to a single event A_i and the number of such events is given by $n = N^{x-1}$. Putting (4) into (7) and replacing the summing over nodes indexes by the summing over the degree distribution $P(k)$ one gets:

$$p_{ij}(x) = 1 - \exp\left[-\frac{k_i k_j}{N} \frac{\langle k(k-1) \rangle^{x-1}}{\langle k \rangle^x}\right] - Q. \quad (8)$$

The assumption underlying (1) is the mutual independence of all contributing events A_i . In fact, since the same edge may participate in several x -walks there exist correlations between these events. Nevertheless, it is easy to see that the fraction of correlated walks is negligible for short walks ($x \ll N$) that play the major role in random graphs showing small-world behavior.

The question is when the term Q in (8) may be neglected. To work out the problem let us perform the following reasoning: if $\forall_{(i,j)}$ there exists $\tilde{\varepsilon} \ll 1$ such that $\tilde{p}_{ij} \leq \tilde{\varepsilon}$ then $\forall_{x \geq 1} \tilde{p}_{iv_1} \tilde{p}_{v_1 v_2} \dots \tilde{p}_{v_{(x-1)}j} | v_{(x-2)} v_{(x-1)} \leq \tilde{\varepsilon}^x \ll 1$ and Q may be ignored. In fact, due to (4) the condition $\tilde{p}_{ij} \ll 1$ is not fulfilled for pairs of vertices i and j possessing large degrees k_i and k_j . The fraction of such pairs may be estimated as

$$\int_{k_{min}}^{k_{max}} P(k_j) \int_{\tilde{\varepsilon}(k)N/k_j}^{k_{max}} P(k_i) dk_i dk_j \ll 1. \quad (9)$$

Using the Chebyshev's inequality [29] and solving (9) with respect to $\tilde{\varepsilon} \ll 1$ one gets the condition when Q may be dropped

$$\frac{\langle k^2 \rangle}{\langle k \rangle^2} (\langle k^2 \rangle - \langle k \rangle^2) \ll N^2. \quad (10)$$

Due to (6) the probability that both vertices are exactly the x -th neighbors may be written as

$$p_{ij}^*(x) = F(x-1) - F(x), \quad (11)$$

where

$$F(x) = \exp\left[-\frac{k_i k_j}{N} \frac{(\langle k^2 \rangle - \langle k \rangle^2)^{x-1}}{\langle k \rangle^x}\right]. \quad (12)$$

Note that averaging (11) over all pairs of vertices one may obtain the intervertex distance distribution $p(x) = \langle \langle p_{ij}^*(x) \rangle \rangle_{i,j}$. Now the mean number of vertices at a certain distance x away from a randomly chosen vertex i can be written as $z_x = \int p_{ij}^*(x) P(k_j) N dk_j$. Taking only the first two terms of power series expansion of both exponential functions in (11) one gets the relationship obtained by Newman et al. [16] $z_x = (z_2/z_1)^{x-1} z_1$ that was received assuming a tree-like structure of random graphs.

The expectation value for the *APL* between i and j is

$$l_{ij}(k_i, k_j) = \sum_{x=1}^{\infty} x p_{ij}^*(x) = \sum_{x=0}^{\infty} F(x). \quad (13)$$

Notice that a walk may cross the same node several times thus the largest possible walk length can be $x = \infty$. The Poisson summation formula allows us to simplify (13)

$$l_{ij}(k_i, k_j) = \frac{-\ln k_i k_j + \ln(\langle k^2 \rangle - \langle k \rangle) + \ln N - \gamma}{\ln(\langle k^2 \rangle / \langle k \rangle - 1)} + \frac{1}{2}, \quad (14)$$

where $\gamma \simeq 0.5772$ is the Euler's constant. The average intervertex distance for the whole network depends on a specified degree distribution $P(k)$

$$l = \frac{\ln(\langle k^2 \rangle - \langle k \rangle) - 2\langle \ln k \rangle + \ln N - \gamma}{\ln(\langle k^2 \rangle / \langle k \rangle - 1)} + \frac{1}{2}. \quad (15)$$

The formulas (14) and (15) diverge when $\langle k^2 \rangle = 2\langle k \rangle$, giving the well-known estimation of percolation threshold in undirected random graphs [32, 33].

To test the formula (15) we start with two well known networks: *ER* classical random graphs and scale-free *BA* networks. The choice of these two networks is not accidental. Both models play an important role in the network science [1, 2, 3, 4]. The *ER* model was historically the first one but it has been realized it is too random to describe real networks. The most striking discrepancy between *ER* model and real networks appears when comparing degree distributions. As mention at the beginning of the paper degree distribution follows power-law in most of real systems, whereas classical random graphs exhibit Poisson degree distribution. The only known mechanism driving real networks into scale-free structures is preferential attachment. The simplest model that incorporates the rule of preferential attachment was originally introduced by Barabási and Albert [9].

Classical ER random graphs. For these networks the degree distribution is given by the Poisson function $P(k) = e^{-\langle k \rangle} \langle k \rangle^k / k!$ and the condition (10) is always fulfilled. However, since $\langle \ln k \rangle$ cannot be calculated analytically for Poisson distribution thus the *APL* may not be directly obtained from (15). To overcome this problem we take advantage of the mean field approximation. Let us assume that all vertices within a graph possess the same degree $\forall_i k_i = \langle k \rangle$. It implies that the *APL* between two arbitrary nodes i and j (15) should describe the average intervertex distance of the whole network

$$l_{ER} = \frac{\ln N - \gamma}{\ln(pN)} + \frac{1}{2}. \quad (16)$$

Until now only a rough estimation of the quantity has been known. One has expected that the average shortest path length of the whole ER graph scales with the number of nodes in the same way as the network diameter. We remind that the diameter d of a graph is defined as the maximal distance between any pair of vertices and $d_{ER} = \ln N / \ln(pN)$ [3, 4]. Fig.1 shows the prediction of the equation (16) in comparison with the numerically calculated *APL* in classical random graphs.

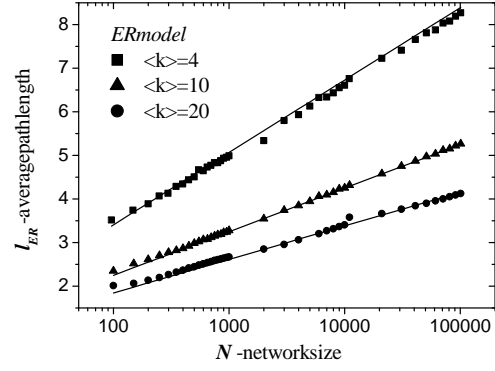


FIG. 1: The average path length l_{ER} versus network size N in *ER* classical random graphs with $\langle k \rangle = pN = 4, 10, 20$. The solid curves represent numerical prediction of Eq.(16).

Scale-free BA networks. The basis of the *BA* model is its construction procedure. Two important ingredients of the procedure are: continuous network growth and preferential attachment. The network starts to grow from an initial cluster of m fully connected vertices. Each new node that is added to the network creates m links that connect it to previously added nodes. The preferential attachment means that the probability of a new link growing out of a vertex i and ending up in a vertex j is given by $\tilde{p}_{ij}^{BA} = mk_j(t_i) / \sum_l k_l(t_i)$, where $k_j(t_i)$ [34] denotes the connectivity of a node j at the time when a new node i is added to the network. Taking into account the time evolution of node degrees in *BA* networks one can show that the probability \tilde{p}_{ij}^{BA} is equivalent to (4). Now let us consider the conditional probability $\tilde{p}_{ij|li}$. Checking the possible time order of the vertices i, j, l it is easy to see that in five of 3! cases $\tilde{p}_{ij|li} = \tilde{p}_{ij}$ and in a good approximation we get instead of (8) the result

$$p_{ij}^{BA}(x) = 1 - \exp \left[-\frac{k_i k_j \langle k^2 \rangle^{x-1}}{N \langle k \rangle^x} \right]. \quad (17)$$

It was found [34] that the degree distribution in *BA* network is given by $P(k) = 2m^2 k^{-\alpha}$, where $k = m, m+1, \dots, m\sqrt{N}$, and the scaling exponent $\alpha = 3$. Putting $\langle k \rangle = 2m$, $\langle k^2 \rangle = m^2 \ln N$ and taking into account (17) one gets that the *APL* between i and j is given by

$$l_{ij}^{BA}(k_i, k_j) = \frac{-\ln(k_i k_j) + \ln N + \ln(2m) - \gamma}{\ln \ln N + \ln(m/2)} + \frac{3}{2}. \quad (18)$$

Averaging (18) over all vertices we obtain

$$l_{BA} = \frac{\ln N - \ln(m/2) - 1 - \gamma}{\ln \ln N + \ln(m/2)} + \frac{3}{2}. \quad (19)$$

Fig.2 shows the *APL* of *BA* networks as a function of the network size N compared with the analytical formula (19). There is a visible discrepancy between the theory and numerical results when $\langle k \rangle = 4$. The discrepancy disappears when the network becomes denser i.e. when

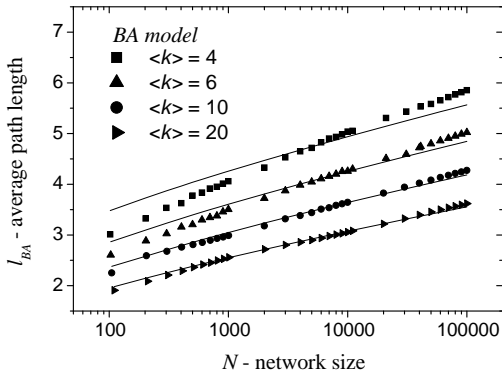


FIG. 2: Characteristic path length l_{BA} versus network size N in BA networks. Solid lines represent Eq.(19).

$\langle k \rangle$ increases. We suspect that it is the effect of structural correlations that occur within the evolving networks [11, 35] and are absent in random graphs used for analytic calculations. The results let us deduce that the correlations become less important in denser networks.

Scale-free networks with arbitrary scaling exponent. Let us consider scale-free random graphs with degree distribution given by a power law, i.e. $P_\alpha(k) = (\alpha - 1)m^{\alpha-1}k^{-\alpha}$, where $k = m, m+1, \dots, mN^{1/(\alpha-1)}$ [17]. Solving (10) for $P_\alpha(k)$ one can see that our approach should work for $\alpha > 2$. Taking advantage of (15) we get that for large networks $N \gg 1$ the APL scales as follows

- $l \simeq 2/(3 - \alpha) + 1/2$ for $2 < \alpha < 3$,
- $l \simeq \ln N / \ln \ln N + 3/2$ for $\alpha = 3$,
- $l \simeq \ln N / (\ln(m(\alpha - 2)/(\alpha - 3) - 1) + 1/2)$ for $\alpha > 3$.

The result for $\alpha \geq 3$ is consistent with estimations obtained by Cohen and Havlin [17]. The first case with l independent on N shows that there is a saturation effect for the mean path length in large networks. Note, that the effect was observed in metabolic networks [36].

In conclusion, we presented a theory for metric properties of random networks with arbitrary degree distribution. The approach is applied to get an analytic formula for the APL in a large class of undirected random graphs with an arbitrary degree distribution $P(k)$. The results are in a very good agreement with numerical simulations performed for ER random graphs and for BA networks. We observed saturation of l in the limit $N \rightarrow \infty$ of scale-free networks with scaling exponents from the range $2 < \alpha < 3$, the small-world behaviour for networks with $\alpha > 3$ and the ultra small-world behaviour of BA model. Our derivations show that the behaviour of APL within scale-free networks is even more intriguing than reported in the recent paper of Cohen and Havlin [17].

Appendix. After finishing the paper we learned about the preprint on this subject written by Dorogovtsev, Mendes and Samukhin [37]. Basing on generating func-

tion formalism the authors derived a similar formula for the APL in random graphs $l \sim \ln N / \ln(\langle k^2 \rangle / \langle k \rangle - 1)$.

Acknowledgments. We are thankful to Sergei Dorogovtsev for critical comments to the preliminary version of this paper. One of us (AF) thanks The State Committee for Scientific Research in Poland for support under grant No. 2P03B01323.

-
- [1] S.Bornholdt and H.G.Schuster, *Handbook of Graphs and networks*, Wiley-Vch (2002).
 - [2] S.N. Dorogovtsev and J.F.F.Mendes, *Evolution of Networks*, Oxford Univ.Press (2003).
 - [3] R.Albert and A.L.Barabási, *Rev. Mod. Phys.* **74** 47 (2002).
 - [4] S.N.Dorogovtsev and J.F.F.Mendes, *Adv.Phys.* **51** 1079 (2002).
 - [5] S.H.Strogatz, *Nature* **410** 268 (2001).
 - [6] S.N.Dorogovtsev et al., *Phys. Rev. Lett.* **85** 4633 (2000).
 - [7] R.Albert and A.L.Barabási, *Phys. Rev. Lett.* **85** 5234 (2000).
 - [8] K.Klemm and V.M.Eguíluz, *Phys. Rev. E* **65** 036123 (2002).
 - [9] A.L.Barabási and R.Albert, *Science* **286**, 509 (1999).
 - [10] P.L.Krapivsky et al., *Phys. Rev. Lett.* **86** 5401 (2001).
 - [11] S.N.Dorogovtsev et al., *cond-mat/0206467* (2002).
 - [12] A.Fronczak et al., *Physica A* **316** 688 (2002).
 - [13] E.Ravasz and A.L.Barabási, *Phys. Rev. E* **67** 026112 (2003).
 - [14] G.Szabó et al., *Phys. Rev. E* **66** 026101 (2002).
 - [15] M.E.J.Newman et al., *Phys. Rev. Lett.* **84** 3201 (2000).
 - [16] M.E.J.Newman et al., *Phys. Rev. E* **64**, 026118 (2001).
 - [17] R.Cohen and S.Havlin, *Phys. Rev. Lett.* **90** 058701 (2003).
 - [18] R.Albert et al., *Nature* **406**, 378 (2000).
 - [19] R.Cohen et al., *Phys. Rev. Lett.* **85** 4626 (2000).
 - [20] R.Cohen et al., *Phys. Rev. Lett.* **86** 3682 (2001).
 - [21] R.Pastor-Satorras et al., *Phys. Rev. Lett* **87** 258701 (2001).
 - [22] V.M.Eguíluz and K.Klemm, *Phys. Rev. Lett.* **89** 108701 (2002).
 - [23] R.Pastor-Satorras and A.Vespignani, *Phys. Rev. Lett.* **86** 3200 (2001).
 - [24] Z.Dezső and A.L.Barabási, *Phys. Rev. E* **65** 055103 (2002).
 - [25] B.J.Kim et al., *Phys. Rev. E* **65** 027103 (2002).
 - [26] L.A.Adamic et al., *Phys. Rev. E* **64** 046135 (2001).
 - [27] S.Valverde et al., *cond-mat/0204344* (2002).
 - [28] D.J.Watts and S.H.Strogatz, *Nature* **393** 440 (1998).
 - [29] W.Feller, *An Introduction to Probability Theory and its Applications*, John Wiley and Sons (1968).
 - [30] R.J.Wilson, *Intr. to Graph Theory*, Longman (1985).
 - [31] S.N.Dorogovtsev et al., *cond-mat/0204111* (2002).
 - [32] D.S.Callaway et al., *Phys. Rev. Lett.* **85** 5468 (2002).
 - [33] A.V.Goltsev et al., *Phys. Rev. E* **67** 026123 (2003).
 - [34] A.L.Barabási et al., *Physica A* **272** 173 (1999).
 - [35] P.L.Krapivsky and S.Redner, *Phys. Rev. E* **63** 066123 (2001).
 - [36] H.Jeong et al., *Nature* **407** 651 (2000).
 - [37] S.N.Dorogovtsev et al., *Nucl. Phys. B* **653** 307 (2003).